

# An Ensemble EM Algorithm for Bayesian Variable Selection

Jin Wang <sup>\*1</sup>, Feng Liang <sup>†1</sup>, and Yuan Ji <sup>‡2</sup>

<sup>1</sup> Department of Statistics, University of Illinois at Urbana-Champaign

<sup>2</sup>Department of Biostatistics, University of Chicago

March 15, 2016

## Abstract

We study the Bayesian approach to variable selection in the context of linear regression. Motivated by a recent work by Ročková and George (2014), we propose an EM algorithm that returns the MAP estimate of the set of relevant variables. Due to its particular updating scheme, our algorithm can be implemented efficiently without inverting a large matrix in each iteration and therefore can scale up with big data. We also show that the MAP estimate returned by our EM algorithm achieves variable selection consistency even when  $p$  diverges with  $n$ . In practice, our algorithm could get stuck with local modes, a common problem with EM algorithms. To address this issue, we propose an ensemble EM algorithm, in which we repeatedly apply the EM algorithm on a subset of the samples with a subset of the covariates, and then aggregate the variable selection results across those bootstrap replicates. Empirical studies have demonstrated the superior performance of the ensemble EM algorithm.

## 1 Introduction

Consider a simple linear regression model with Gaussian noise:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{1}$$

where  $\mathbf{y} = (y_1, \dots, y_n)^T$  is the  $n \times 1$  response,  $\mathbf{e} = (e_1, \dots, e_n)^T$  is a vector of iid Gaussian random variables with mean 0 and variance  $\sigma^2$ , and  $\mathbf{X}$  is the  $n \times p$  design matrix. The

---

<sup>\*</sup>jinwang8@illinois.edu

<sup>†</sup>liangf@illinois.edu

<sup>‡</sup>jiyuan@uchicago.edu

unknown parameters are the regression parameter  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  and the error variance  $\sigma^2$ . In many real applications such as bioinformatics and image analysis, where linear regression models have been routinely used, the number of potential predictors (i.e.,  $p$ ) is large but only a small fraction of them is believed to be relevant. Therefore the linear model (1) is often assumed to be “sparse” in the sense that most of the coefficients  $\beta_j$ ’s are zero. Estimating the set of relevant variables,  $S = \{j : \beta_j \neq 0\}$ , is an important problem in modern statistical analysis.

The Bayesian approach to variable selection is conceptually simple and straightforward. First introduce a  $p$ -dimensional binary vector  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)^T$  to index all the  $2^p$  sub-models, where  $\gamma_j = 1$  if the  $j$ th variable is included in this model and 0 if excluded. Usually  $\gamma_j$ ’s are modeled by independent Bernoulli distributions. Given  $\boldsymbol{\gamma}$ , a popular prior choice for  $\boldsymbol{\beta}$  is the “spike and slab” prior (Mitchell and Beauchamp, 1988):

$$\pi(\beta_j \mid \gamma_j) = \begin{cases} \delta_0(\beta_j), & \text{if } \gamma_j = 0; \\ g(\beta_j), & \text{if } \gamma_j = 1, \end{cases} \quad (2)$$

where  $\delta_0(\cdot)$  is the Kronecker delta function corresponding to the density function of a point mass at 0 and  $g$  is a continuous density function. After specifying priors on all the unknowns, one needs to calculate the posterior distribution. Most algorithms for Bayesian variable selection rely on MCMC such as Gibbs or Metropolis Hasting to obtain the posterior distribution; for a review on recent developments in this area, see O’Hara and Sillanpää (2009). MCMC algorithms, however, are insufficient to meet the growing demand on scalability from real applications. Since the primary goal is variable selection, we focus on efficient algorithms that return the MAP estimate of  $\boldsymbol{\gamma}$ , as an alternative to these MCMC-based sampling methods that return the whole posterior distribution on all the unknown parameters.

Recently, Ročková and George (2014) proposed a simple, elegant EM algorithm for Bayesian variable selection. They adopted a continuous version of the “spike and slab” prior—the spike component in (2) is replaced by a normal distribution with a small variance (George and McCulloch, 1993), and proposed an EM algorithm to obtain the MAP estimate of the regression coefficient  $\boldsymbol{\beta}$ . The MAP estimate  $\hat{\boldsymbol{\beta}}_{\text{MAP}}$ , however, is not sparse, and an additional thresholding step is needed to estimate  $\boldsymbol{\gamma}$ .

In this paper, we develop an EM algorithm that directly returns the MAP estimate of  $\boldsymbol{\gamma}$ , so no further thresholding is needed. We adopt the same continuous “spike and slab” prior. Different from the algorithm by Ročková and George (2014) that returns

$\hat{\beta}_{\text{MAP}}$  by treating  $\gamma$  as latent, our algorithm returns the MAP estimate of the model index,  $\hat{\gamma}_{\text{MAP}}$ , by treating  $\beta$  as latent. The special structure of our EM algorithm allows us to use a computational trick to avoid inverting a big matrix at each iteration, which seems unavoidable in the algorithm by Ročková and George (2014). Further we can show that the  $\hat{\gamma}_{\text{MAP}}$  achieves asymptotic consistency even when  $p$  diverges to infinity with the sample size  $n$ .

Although shown to achieve selection consistency, in practice, our EM algorithm could get stuck at a local mode due to the large discrete space in which  $\gamma$  lies. Borrowing the idea of bagging, we propose an ensemble version of our EM algorithm (which we call BBEM): apply the algorithm on multiple Bayesian bootstrap (BB) copies of the data, and then aggregate the variable selection results. Bayesian bootstrap for variable selection was explored before by Clyde and Lee (2001) for the purpose of prediction, where models built on different bootstrap copies are combined to predict the response. But the focus of our approach is to summarize the evidence for variable relevance from multiple BB copies, which is similar in nature to several frequentist ensemble methods for variable selection, such as the AIC ensemble (Zhu and Chipman, 2006), stability selection (Meinshausen and Bühlmann, 2010), and random Lasso (Wang et al., 2011).

The remaining of the paper is organized as follows. Section 2 describes the EM algorithm in detail, Section 3 presents the asymptotic results, and Section 4 describes the BBEM algorithm. Empirical studies are presented in Section 5 and conclusions and remarks in Section 6.

## 2 The EM Algorithm

### 2.1 Prior Specification

We adopt the continuous version of “spike and slab” prior for  $\beta$ , i.e. a mixture of two normal components with mean zero and different variances:

$$\pi(\beta_j \mid \sigma, \gamma_j) = \begin{cases} \text{N}(0, \sigma^2 v_0), & \text{if } \gamma_j = 0; \\ \text{N}(0, \sigma^2 v_1), & \text{if } \gamma_j = 1, \end{cases} \quad (3)$$

where  $v_1 > v_0 > 0$ . Alternatively, we can write the prior on  $\beta$  as

$$\pi(\beta_j \mid \sigma^2, \gamma_j) = \text{N}(0, \sigma^2 d_{\gamma_j}),$$

where

$$d_{\gamma_j} = \gamma_j v_1 + (1 - \gamma_j) v_0.$$

For the remaining parameters, we specify independent Bernoulli priors on elements of  $\gamma$ , and conjugate priors like Beta and Inverse Gamma on  $\theta$  and  $\sigma^2$ , respectively:

$$\begin{aligned}\pi(\gamma \mid \theta) &= \text{Bern}(\theta), \\ \pi(\theta) &= \text{Beta}(a_0, b_0), \\ \pi(\sigma^2) &= \text{IG}(\nu/2, \nu\lambda/2).\end{aligned}$$

For hyper-parameters  $(a_0, b_0, \nu, \lambda)$ , we suggest the following non-informative choices unless prior knowledge is available:

$$a_0 = b_0 = 1.1, \quad \nu = \lambda = 1. \quad (4)$$

The choice for  $v_0$  and  $v_1$  will be discussed later.

## 2.2 The Algorithm

With the Gaussian model and prior distributions specified above, we can write down the full posterior distribution:

$$\pi(\gamma, \beta, \theta, \sigma^2 \mid \mathbf{y}) \propto p(\mathbf{y} \mid \beta, \sigma^2) \times \pi(\beta \mid \sigma, \gamma) \times \pi(\gamma \mid \theta) \times \pi(\theta) \times \pi(\sigma^2).$$

Treating  $\beta$  as the latent variable, we derive an EM algorithm that returns the MAP estimation of parameters  $\Theta = (\gamma, \sigma^2, \theta)$ , whereas the roles of  $\beta$  and  $\gamma$  are switched in Ročková and George (2014).

### E Step

The objective function  $Q$  at the  $(t+1)$ -th iteration in an EM algorithm is defined as the integrated logarithm of the full posterior with respect to  $\beta$  given  $\mathbf{y}$  and the parameter values from the previous iteration  $\Theta^{(t)} = (\gamma^{(t)}, \sigma_{(t)}^2, \theta^{(t)})$ , i.e.,

$$\begin{aligned}Q(\Theta \mid \Theta^{(t)}) &= \mathbb{E}_{\beta \mid \Theta^{(t)}, \mathbf{y}} \log \pi(\Theta, \beta \mid \mathbf{y}) \\ &= -\frac{1}{2\sigma^2} \mathbb{E}_{\beta \mid \Theta^{(t)}, \mathbf{y}} [\|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^p \frac{\beta_j^2}{d_{\gamma_j}}] + F(\Theta),\end{aligned} \quad (5)$$

where

$$\begin{aligned}F(\Theta) &= -\frac{n+p}{2} \log \sigma^2 - \frac{1}{2} \sum_{j=1}^p \log d_{\gamma_j} + \pi(\gamma \mid \theta) \\ &\quad + \log \pi(\theta) + \log \pi(\sigma^2) + \text{Constant}\end{aligned}$$

is a function of  $\Theta$  not depending on  $\beta$ .

It is easy to show that  $\beta$  follows a Normal distribution with mean  $\mathbf{m}$  and covariance matrix  $\sigma_{(t)}^2 \mathbf{V}$ , given  $\Theta^{(t)}$  and  $\mathbf{y}$ , where

$$\begin{aligned} \mathbf{m} &= \mathbf{V}^{-1} \mathbf{X}^T \mathbf{y}, \quad \mathbf{V} = (\mathbf{X}^T \mathbf{X} + D_{\gamma^{(t)}}^{-1})^{-1}, \\ D_{\gamma^{(t)}} &= \text{diag}\left(d_{\gamma_j^{(t)}}\right)_{j=1}^p = \text{diag}\left(\gamma_j^{(t)} v_1 + (1 - \gamma_j^{(t)}) v_0\right)_{j=1}^p. \end{aligned} \quad (6)$$

Then the two expectation terms in (5) can be expressed as:

$$\mathbb{E}_{\beta|\Theta^{(t)}, \mathbf{y}} \|\mathbf{y} - \mathbf{X}\beta\|^2 = \sigma_{(t)}^2 \text{tr}(\mathbf{X}\mathbf{V}\mathbf{X}^T) + \|\mathbf{y} - \mathbf{X}\mathbf{m}\|^2, \quad (7)$$

$$\mathbb{E}_{\beta|\Theta^{(t)}, \mathbf{y}} \sum_{j=1}^p \frac{\beta_j^2}{d_{\gamma_j}} = \sum_{j=1}^p \frac{\sigma_{(t)}^2 V_{jj} + m_j^2}{(1 - \gamma_j^{(t)}) v_0 + \gamma_j^{(t)} v_1}. \quad (8)$$

## M Step

We sequentially update parameters  $(\gamma, \theta, \sigma)$  to maximize the objective function  $Q$ .

1. **Update  $\gamma_j$ 's.** The terms involving  $\gamma_j$  in (5) are

$$-\frac{1}{2\sigma_{(t)}^2} \mathbb{E}_{\beta|\Theta^{(t)}, \mathbf{y}} \left[ \frac{\beta_j^2}{d_{\gamma_j}} \right] - \frac{1}{2} \log d_{\gamma_j} + \log \pi(\gamma_j | \theta^{(t)}). \quad (9)$$

Plug in  $\gamma_j = 0$  and  $\gamma_j = 1$  to (9) respectively, then we have

$$\gamma_j^{(t+1)} = 1, \quad \text{if } \mathbb{E}_{\beta|\Theta^{(t)}, \mathbf{y}} [\beta_j^2] > r^{(t)}, \quad (10)$$

where

$$r^{(t)} = \frac{\sigma_{(t)}^2}{1/v_0 - 1/v_1} \left( \log \frac{v_1}{v_0} - 2 \log \frac{\theta^{(t)}}{1 - \theta^{(t)}} \right).$$

2. **Update  $(\sigma^2, \theta)$ .** Given  $\gamma^{(t+1)}$ , the updating equations for the other two parameters are given by

$$\sigma_{(t+1)}^2 = \frac{\mathbb{E}_{\beta|\Theta^{(t)}, \mathbf{y}} \left[ \|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^p \beta_j^2 / d_{\gamma_j^{(t+1)}} \right] + \nu \lambda}{n + p + \nu}, \quad (11)$$

$$\theta^{(t+1)} = \frac{\sum_{j=1}^p \gamma_j^{(t+1)} + a_0 - 1}{p + a_0 + b_0 - 2}. \quad (12)$$

## Stopping Rule

The EM algorithm alternates between the E-step and M-step until convergence. A natural stopping criterion is to check whether the change of the objective function  $Q$  is small. To reduce the computation cost for evaluating the  $Q$  function, we adopt a different stopping rule as our main focus is  $\gamma$ : we stop our algorithm when the estimate  $\gamma^{(t)}$  stays the same for  $k_0$  iterations. In practice, we suggest to set  $k_0 = 3$ . The pseudo code of this EM algorithm is summarized in Algorithm 1.

### Algorithm 1: EM Algorithm

**Input:**  $\mathbf{X}, \mathbf{y}, v_0, v_1, a_0, b_0, \nu, \lambda$

Initialize  $\Theta^{(0)}$ ;

E-step: Calculate the two expectations in (7) and (8), denoted as  $EE^{(0)}$ ;

**for**  $t = 1 : \text{maxIter}$  **do**

    M-step: Update  $\Theta^{(t)}$  from Eq (10, 11, 12);

    E-step: Update  $EE^{(t)}$  from Eq (7, 8);

**if**  $\gamma^{(t)}$  stays the same for  $k_0 = 3$  iterations **then**

        break;

**end**

**end**

Return  $\gamma, \mathbf{m}$ ;

## 2.3 Computation Cost

At each E-step, updating the posterior of  $\beta$  given other parameters in (6) requires inverting a  $p \times p$  matrix

$$\mathbf{V}_{(t)} = (\mathbf{X}^T \mathbf{X} + D_{\gamma^{(t)}}^{-1})^{-1}, \quad (13)$$

which is the major computational burden of this algorithm. When  $p > n$ , we can use the Sherman-Morrison-Woodbury formula to compute the inverse of an  $n \times n$  matrix. So the computation cost at each iteration is of order  $O(\min(n, p)^3)$ . It is, however, still time-consuming when both  $n$  and  $p$  are large.

Note that the only thing that changes in (13) from iteration to iteration is  $D_{\gamma^{(t)}}$ , a diagonal matrix depending on the binary vector  $\gamma^{(t)}$ . From our experience, only a small fraction of  $\gamma_j^{(t)}$ 's are changed at each iteration after the first a couple of iterations. So the

idea is to use the following recursive formula to compute  $\mathbf{V}_{(t)}$ :

$$\begin{aligned}\mathbf{V}_{(t)} &= (\mathbf{X}^T \mathbf{X} + D_{\boldsymbol{\gamma}^{(t-1)}}^{-1} + D_{\boldsymbol{\gamma}^{(t)}}^{-1} - D_{\boldsymbol{\gamma}^{(t-1)}}^{-1})^{-1} \\ &= (\mathbf{V}_{(t-1)}^{-1} + D_{\boldsymbol{\gamma}^{(t)}}^{-1} - D_{\boldsymbol{\gamma}^{(t-1)}}^{-1})^{-1}\end{aligned}\quad (14)$$

where  $D_{\boldsymbol{\gamma}^{(t)}}^{-1} - D_{\boldsymbol{\gamma}^{(t-1)}}^{-1}$  is a diagonal matrix with the  $j$ -th diagonal entry being non-zero only if the inclusion/exclusion status, i.e., the value of  $\gamma_j$ , is changed from the last iteration. Let  $l$  denote the number of variables whose  $\gamma_j$  values are changed from iteration  $(t-1)$  to  $t$ . Then  $D_{\boldsymbol{\gamma}^{(t)}}^{-1} - D_{\boldsymbol{\gamma}^{(t-1)}}^{-1}$  is a rank  $l$  matrix. We can apply the Woodbury formula on (14) to reduce the computation complexity from  $O(\min(n, p)^3)$  to  $O(l^3)$ .

For example, without loss of generality, suppose only the first  $l$  covariates have their  $\gamma_j$  values changed. Then, we can write

$$D_{\boldsymbol{\gamma}^{(t)}}^{-1} - D_{\boldsymbol{\gamma}^{(t-1)}}^{-1} = U_{p \times l} A_{l \times l} U^T,$$

where  $A = (\frac{1}{v_0} - \frac{1}{v_1}) \text{diag}(2\gamma_j^{(t)} - 1)_{j=1}^l$  and  $U$  consists of the first  $l$  columns from  $\mathbf{I}_p$ . Applying the Woodbury formula, we have

$$\mathbf{V}_{(t)} = \mathbf{V}_{(t-1)} - \mathbf{V}_{(t-1)} U (A^{-1} + U^T \mathbf{V}_{(t-1)} U)^{-1} U^T \mathbf{V}_{(t-1)}.$$

### 3 Asympototic Consistency

In this section, we study the asymptotic property of  $\hat{\boldsymbol{\gamma}}_n$ , the MAP estimate of model index returned by our EM algorithm. Assume the data  $\mathbf{y}_n$  are generated from a Gaussian regression model:

$$\mathbf{y}_n \sim \mathbf{N}_n(\mathbf{X}_n \boldsymbol{\beta}_n^*, \sigma^2 \mathbf{I}_n).$$

Here we consider a triangular array set up: the dimension  $p = p_n$  diverges with  $n$  and the true coefficients  $\boldsymbol{\beta}_n^*$  also vary with  $n$ . Suppose the true model is indexed by  $\boldsymbol{\gamma}_n^*$ , where  $\gamma_{nj}^* = 1$  if  $\beta_{nj}^* \neq 0$  and  $\gamma_{nj}^* = 0$  if  $\beta_{nj}^* = 0$ . We show that our EM algorithm has the following selection consistency property:

$$\mathbb{P}(\hat{\boldsymbol{\gamma}}_n = \boldsymbol{\gamma}_n^*) \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

First we list some regularity conditions needed in our proof. Let  $\lambda_{\min}(A)$  denote the

smallest eigenvalue of matrix  $A$ . We assume

- (A1)  $\lambda_{\min}(\mathbf{X}_n^T \mathbf{X}_n)^{-1} = O(n^{-\eta_1})$ ,  $0 < \eta_1 \leq 1$ ;
- (A2)  $\|\beta_n^*\|_2 = O(n^{\eta_2})$ ,  $0 < \eta_2 < \eta_1$ ;
- (A3)  $\liminf_n \frac{\min\{|\beta_{nj}^*|, \gamma_{nj}^* = 1\}}{n^{(\eta_3-1)/2}} \geq M$ ,  $0 \leq \eta_3 < 1$ ;
- (A4)  $a_0 \sim p_n$ ,  $b_0 \sim p_n$ ,  $\nu = \infty$ ,  $\lambda = 1$ ,

where  $M$  is a positive constant, and  $(a_0, b_0, \nu, \lambda)$  are the hyper-parameters from the Beta and InvGamma priors.

Assumption (A1) controls the collinearity among covariates; in the traditional asymptotic setting where  $p$  is fixed, we have  $\eta_1 = 1$ . Assumption (A2) controls the sparsity (in terms of  $L_2$  norm) of the true regression coefficient vector. Assumption (A3) requires that the minimal non-zero coefficient cannot go to zero at a rate faster than  $1/\sqrt{n}$ ; in the traditional asymptotic setting where  $\beta^*$  is fixed, we have  $\eta_3 = 0$ . Assumption (A4) is purely technical, which ensures that  $\hat{\theta}_n$  and  $\hat{\sigma}_n^2$  are bounded. In fact we could fix  $\hat{\theta}_n$  and  $\hat{\sigma}_n^2$  to be any constant, which does not affect the proof. In our simulation studies, we still recommend (4) as the choice for hyper-parameters unless  $p$  is large.

**Theorem 3.1.** *Assume (A1-A4) and  $p = O(n^\alpha)$  where  $0 \leq \alpha < 1$ . With  $v_1$  fixed and  $v_0$  satisfying*

$$0 < v_0 = O(n^{-r_0}), \quad 1 - \eta_3 < r_0 < \min\left\{\eta_1 - \alpha, \frac{2}{3}(\eta_1 - \eta_2)\right\},$$

*the model returned by our EM algorithm,  $\hat{\gamma}_n$ , achieves the following selection consistency,*

$$\mathbb{P}(\hat{\gamma}_n = \gamma_n^*) \rightarrow 1, \quad \text{as } n \rightarrow \infty. \quad (15)$$

*Proof.* See Appendix. □

## 4 The BBEM Algorithm

A common issue with EM algorithms is that they could be trapped at a local maximum. There are some standard remedies available for dealing with this issue, for instance, trying a set of different initial values or utilizing some more advanced optimization procedures at the M-step. Since our EM algorithm is searching for the optimal  $\gamma$  over a big discrete space, all  $p$ -dimensional binary vectors, these remedies are less useful when  $p$  is large.

When doing optimization with  $\gamma$ , a discrete vector, the resulting solution is often not stable, i.e., has a large variance. Bagging is an easy but powerful method (Breiman, 1996)



for variance reduction, which applies the same algorithm on multiple bootstrap copies of the data, and then aggregates the results. We proposed the following ensemble EM algorithm, in which we repeatedly run the EM variable selection algorithm, Algorithm 1 from Section 2.2, on Bayesian bootstrap replicates.

The original bootstrap repeatedly draws samples from the original data set  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$  with replacement, i.e., each observation  $(\mathbf{x}_i, y_i)$  is sampled with probability  $1/n$ . In Bayesian bootstrap (Rubin, 1981), instead of sampling a subset of the data, we assign a random weight  $w_i$  to the  $i$ -th observation and then fit a weighted least squares regression model on the whole data set. In particular, following Rubin (1981), we generate the weights  $\mathbf{w} = (w_1, \dots, w_n)$  from a  $n$ -category Dirichlet distribution:

$$\mathbf{w}_{n \times 1} \sim \text{Dir}(1, \dots, 1). \quad (16)$$

When applying Algorithm 1 on a weighted linear regression model, all the updating equations stay the same, except equation (6) for the posterior of  $\beta$ , which should be changed to:

$$\mathbf{m} = \mathbf{V} \mathbf{X}^T \text{diag}(\mathbf{w}) \mathbf{y}, \quad \mathbf{V} = (\mathbf{X}^T \text{diag}(\mathbf{w}) \mathbf{X} + D_{\gamma^{(t)}}^{-1})^{-1}. \quad (17)$$

Eq (7), the expectation of the weighted residual sum of squares, should also be changed accordingly:

$$\mathbb{E}_{\beta|\Theta^{(t)}, \mathbf{y}} \|\mathbf{y} - \mathbf{X}\beta\|_{\mathbf{w}}^2 = \sigma_{(t)}^2 \text{tr}(\text{diag}(\mathbf{w}) \mathbf{X} \mathbf{V} \mathbf{X}^T) + (\mathbf{y} - \mathbf{X}\mathbf{m})^T \text{diag}(\mathbf{w}) (\mathbf{y} - \mathbf{X}\mathbf{m}). \quad (18)$$

It is well-known that in order to make the aggregation work, we should control the correlation among estimates from bootstrap replicates. For example, in random forest (Breiman, 2001), the number of variables used for choosing the optimal split of a tree is restricted to a subset of the variables, instead of using all  $p$  variables. A similar idea was implemented in Random Lasso (Wang et al., 2011), an ensemble algorithm for variable selection. In the same spirit, we apply the EM algorithm only on a subset of the variables at each Bayesian bootstrap iteration. A naive way is to randomly pick a subset from the  $p$  variables. This, however, will be inefficient when  $p$  is large and the true model is sparse, since it is likely most random subsets will not contain any relevant variables. So we employ a biased sampling procedure: sample the  $p$  variables based on a weight vector  $\tilde{\pi}$  that is defined as

$$\tilde{\pi}_{p \times 1} \propto |\mathbf{X}^T \mathbf{y}| / \text{diag}(\mathbf{X}^T \mathbf{X}), \quad (19)$$

that is, variables are sampled based on their marginal effect in a simple linear regression.

The ensemble EM algorithm operates as follows. First we sample a random set of  $L$  variables according to the probability vector  $\tilde{\pi}$ , and draw a  $n \times 1$  bootstrap weight vector  $\mathbf{w}$  from (16). Let  $\tilde{\mathbf{X}}$  be the new data matrix with the  $L$  columns. Then apply the EM algorithm on  $\tilde{\mathbf{X}}$  with weight  $\mathbf{w}$ . Let  $\gamma_k$  denote the model returned by the  $k$ -th Bayesian bootstrap iteration, where the  $j$ -th element of  $\gamma_k$  is 1 if the  $j$ -th variable is selected and zero otherwise; of course, the  $j$ -th element is zero if the  $j$ -th variable is not included in the initial  $L$  variables. Define the final variable selection frequency for the  $p$  variables as

$$\phi_{p \times 1} = \frac{1}{K} \sum_{k=1}^K \gamma_k. \quad (20)$$

We can report the final variable selection result by thresholding  $\phi_j$ 's at some fixed number, for example, a half. Or we can produce a path-plot of  $\phi$  as  $v_0$  varies, which could be a useful tool to investigate the importance of each variable. We illustrate this in our simulation study in Section 5.

As for the computational cost, the inversion of the  $L \times L$  matrix in (17) is a big improvement compared with that of a  $p \times p$  matrix, while it can be further simplified through the fast computing trick in Section 2.3. We call this algorithm, BBEM, which is summarized in Algorithm 2.

## 5 Empirical Study

In this section, we first compare the proposed EM algorithm (Algorithm 1) with other popular methods on a widely used benchmark data set. Then we compare BBEM (Algorithm 2) with other methods on two more challenging data sets of larger dimensions. Finally, we applied BBEM on a restaurant revenue data from a Kaggle competition, and showed that our algorithm outperforms the benchmark from random forest.

For the hyper-parameters  $v_0$  and  $v_1$ , we set  $v_1 = 100$  as fixed and tune an appropriate value for  $v_0$  either based on 5-fold cross-validation or BIC. For the initial value for  $\theta$ , we suggest to use  $1/2$  for ordinary problems, but  $\sqrt{n}/p$  for large- $p$  problems. The initial value of  $\sigma^2$  is set as 1. In addition, there are two bootstrap parameters: the total number of replicates  $K$  and the number of variables used in each bootstrap  $L$ . For efficiency, the number of variables in each bootstrap replicate should not exceed the sample size  $n$ . We use  $K = 100$ , and  $L = n/2 = 50$  if  $p$  is large and  $L = p$  if  $p$  is small.

**Algorithm 2:** BBEM Algorithm**Input:**  $\mathbf{X}, \mathbf{y}, v_0, v_1, a_0, b_0, \nu, \lambda, K, L$ Compute the variable weight  $\tilde{\pi}$  from (19);**for**  $k = 1 : K$  **do**    Generate a subset of  $L$  variables according to  $\tilde{\pi}$ ;    Make the replicate  $\tilde{\mathbf{X}}^k$  with the  $L$  variables;    Initialize  $\Theta_k^{(0)}$ ;    Generate bootstrap weight  $\mathbf{w}$  from (16);    E-step: Calculate the two expectations in (8), denoted as  $EE_k^{(0)}$ ;    **for**  $t = 1 : \text{maxIter}$  **do**        M-step: Update  $\Theta_k^{(t)}$  from Eq (10, 11, 12);        E-step: Update  $EE_k^{(t)}$  from Eq (18, 8);        **if**  $\gamma_k^{(t)}$  *stays the same for*  $k_0 = 3$  *iterations* **then**

break;

**end**    **end**    Record  $\gamma_k^{(t)}, \mathbf{m}_k^{(t)}$ ;**end**Return  $\phi$  from Eq (20);

## 5.1 A widely used benchmark

First we apply our EM algorithm on a widely used benchmark data set (Tibshirani, 1996), which has  $p = 8$  variables, each from a standard normal distribution with pairwise correlation  $\rho(\mathbf{x}_i, \mathbf{x}_j) = 0.5^{|i-j|}$ . The response variable is generated from

$$\mathbf{y} = 3\mathbf{x}_1 + 1.5\mathbf{x}_2 + 2\mathbf{x}_5 + \epsilon$$

where  $\epsilon \sim N(0, \sigma^2)$ .

Following Fan and Li (2001), we repeat the experiment 100 times under two scenarios: (1)  $n = 40, \sigma = 3$  and (2)  $n = 60, \sigma = 1$ . The result is shown in Table 1, which reports the average number of zero-coefficients (i.e., no selection) among signal variables ( $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_5$ ) and among noise variables, respectively. The results for SCAD1 (tuning parameter selected by cross-validation), SCAD2 (tuning parameter fixed) and LASSO are taken from Fan and Li (2001). In the first “small sample-size high noise” scenario, our EM algorithm has the highest number of zero-coefficients among noise variables, i.e., the lowest type I error. The average number of signal variables missed by EM is slightly higher than SCAD1 (where the tuning parameter is chosen by cross-validation) but less than SCAD2 (where the tuning parameter is pre-fixed). But overall, our EM algorithm and the two SCAD methods perform the best. In the second “large sample-size low noise” scenario, no signal variables are missed by any method, but EM has the lowest type I error.

Following Wang et al. (2011) and Xin and Zhu (2012), we repeat the experiment 100 times with the same sample size  $n = 50$  but two different noise levels: low noise level ( $\sigma = 3$ ) and high noise level ( $\sigma = 6$ ). Table 2 reports the minimum, median, maximum of being selected out of 100 simulations for the signal and the noise variables, respectively. Both Lasso and random Lasso have a higher chance of selecting the signal variables, but at the price of mistakenly including many noise variables. Overall, our EM algorithm performs the best, along with PGA and stability selection, two frequentist ensemble methods for variable selection.

## 5.2 A highly-correlated data

Next we demonstrate our two algorithms on a highly-correlated example from Wang et al. (2011). The data has  $p = 40$  variables and the response  $\mathbf{y}$  is generated from

$$\mathbf{y} = 3\mathbf{x}_1 + 3\mathbf{x}_2 - 2\mathbf{x}_3 + 3\mathbf{x}_4 + 3\mathbf{x}_5 - 2\mathbf{x}_6 + \epsilon,$$

Method	$\mathbf{x}_j \in \text{Noise}$ ( $j=3,4,6,7,8$ )	$\mathbf{x}_j \in \text{Signal}$ ( $j=1,2,5$ )
$n = 40, \sigma = 3$		
EM	4.55	0.24
SCAD1	4.20	0.21
SCAD2	4.31	0.27
LASSO	3.53	0.07
Oracle	5.00	0.00
$n = 60, \sigma = 1$		
EM	4.72	0.00
SCAD1	4.37	0.00
SCAD2	4.42	0.00
LASSO	3.56	0.00
Oracle	5.00	0.00

Table 1: A widely used benchmark. The average number of zero-coefficients (i.e., no selection) out of 100 simulations for each types of variable (Signal or Noise) are shown. The results other than EM (Alg 1) are from Fan and Li (2001).

where  $\epsilon \sim N(0, \sigma^2)$  and  $\sigma = 6$ . Each  $\mathbf{x}_i$  is generated from a standard normal with the following correlation structure among the first six signal variables: the signal variables are divided into two groups,  $V_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$  and  $V_2 = \{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$ ; the within group correlation is 0.9 and the between-group correlation is 0.

We repeat the simulation 100 times with  $n = 50$  and  $n = 100$ , and the results are summarized in Table 3. For this example, due to the high correlation among features we expect ensemble methods to perform better. Indeed, BBEM has the best performance in terms of selecting true signal variables while controlling the error of including noise variables. The performance of the EM algorithm, although not the best, is also comparable with other top ensemble methods like random Lasso from Wang et al. (2011), and T2E and PGA from Xin and Zhu (2012).

For illustration purpose, we apply BBEM on a data set with  $n = 50$  and  $v_0$  varying from  $10^{-4}$  to 1. Figure 1 shows the path-plot of the selection frequency from BBEM. There is clearly a gap between the signal variables and the noise ones. For a range of  $v_0$ , from 0.001 to 0.02, BBEM can successfully select the six true variables  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_6\}$  if we threshold the selection frequency  $\phi_j$  at 0.5.

Method	$\mathbf{x}_j \in \text{Signal } (j=1,2,5)$			$\mathbf{x}_j \in \text{Noise } (j=3,4,6,7,8)$		
	Min	Median	Max	Min	Median	Max
$n = 50, \sigma = 3$						
EM	91	97	100	3	6	12
Lasso	99	100	100	48	55	61
Random Lasso	95	99	100	33	40	48
ST2E	89	96	100	4	12	20
PGA	82	98	100	4	7	11
Stability selection						
$\lambda_{min} = 1$	81	83	100	0	2	9
$\lambda_{min} = 0.5$	90	98	100	4	8	22
$n = 50, \sigma = 6$						
EM	53	67	91	6	10	14
Lasso	76	85	99	47	49	53
Random Lasso	92	94	100	40	48	58
ST2E	68	69	96	9	13	21
PGA	54	76	94	9	14	16
Stability selection						
$\lambda_{min} = 1$	59	61	92	4	8	18
$\lambda_{min} = 0.5$	76	84	100	30	42	50

Table 2: A widely used benchmark. The min, median, max number of being selected out of 100 simulations for each types of variable (Signal or Noise) are shown. The results other than EM (Alg 1) are from Xin and Zhu (2012).

### 5.3 A Large- $p$ small- $n$ example

Finally we apply BBEM on a large- $p$  small- $n$  example from Ročková and George (2014), where  $p = 1000$  and  $n = 100$ . Each of the  $p$  features is generated from a standard normal with pairwise correlation to be  $0.6^{|i-j|}$  and the response  $\mathbf{y}$  is generated from the following linear model:

$$\mathbf{y} = \mathbf{x}_1 + 2\mathbf{x}_2 + 3\mathbf{x}_3 + \epsilon,$$

where  $\epsilon \sim N(0, 3)$ .

For this large  $p$  example, we set the parameters in the BBEM algorithm as follows: the initial value of  $\theta$  is  $\sqrt{n}/p$ , the number of variables used in each bootstrap iteration

Method	$\mathbf{x}_j \in \text{Signal } (j= 1:6)$			$\mathbf{x}_j \in \text{Noise}$		
	Min	Median	Max	Min	Median	Max
$n = 50, \sigma = 6$						
Lasso	11	70	77	12	17	25
Random Lasso	84	96	97	11	21	30
ST2E	85	96	100	18	25	34
PGA	55	87	90	14	23	32
EM	65	85.5	89	4	10	13
BBEM	89	96	100	4	8	15
$n = 100, \sigma = 6$						
Lasso	8	84	88	12	22	31
Random Lasso	89	99	99	8	14	21
ST2E	93	100	100	14	21	27
PGA	40	85	92	13	22	33
EM	84	91	95	1	7	16
BBEM	95	99	100	4	9	14

Table 3: A highly-correlated data. The min, median, max number of times being selected (i.e., no selection) out of 100 simulations for each type of variables (Signal and Noise) are shown. The results other than EM and BBEM are from Xin and Zhu (2012).

$L = n/2 = 50$  and the total number of replicates  $K = 100$ . It is well known that cross-validation based on prediction accuracy tends to include more noise variables. So, for this example where the true model is known to be sparse, we choose to tune  $v_0$  via BIC. For illustration purpose, we also include BBEM with a fixed tuning parameter  $v_0 = 0.03$  in the comparison group. We compare BBEM with the EMVS algorithm from Ročková and George (2014), which is implemented by us using the annealing technique for  $\beta$ 's initialization, and fixed  $v_0 = 0.5, v_1 = 1000$  as suggested in Ročková and George (2014).

Table 5.3 reports the average number of signal and noise variables being selected over 100 iterations for each method. BBEM with BIC tuning performs the best: it selects 2.99 signal variables out of 3 on average (i.e., only miss one variable, the weakest signal  $\mathbf{x}_1$ , once in all 100 iterations) and meanwhile has the smallest type I error. The BBEM algorithm with a fixed tuning parameter has a similar result as EMVS but is much faster. The computation advantage for BBEM comes from two aspects: the computation trick that reduces the computation cost on matrix inversion and the sub-sampling step in Bayesian

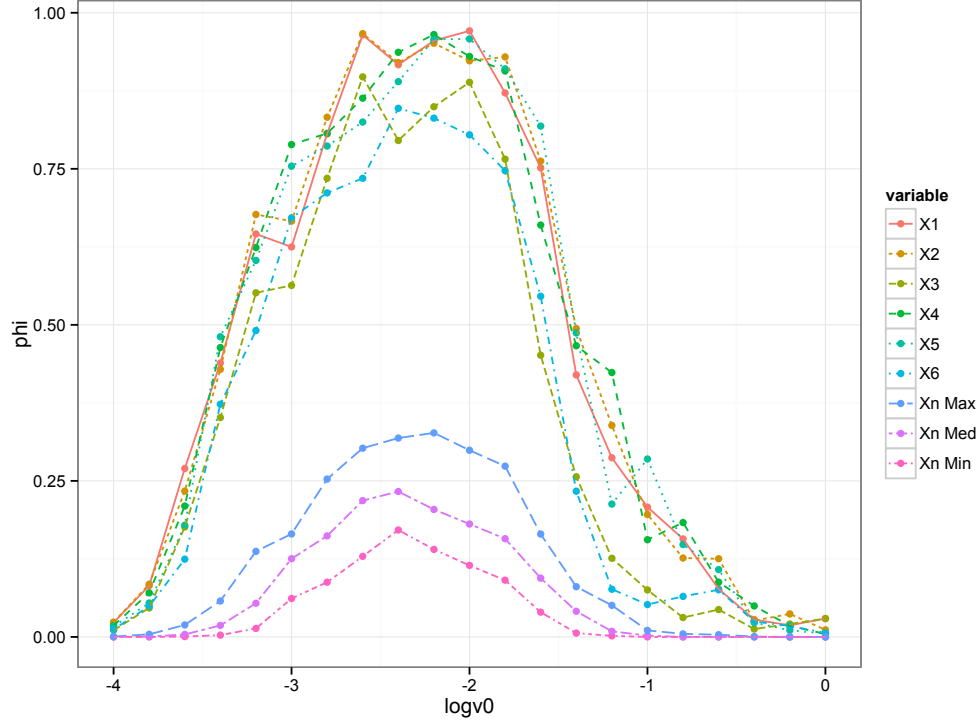


Figure 1: Highly-correlated data  $n = 50$ . A path-plot of the average selection frequency when  $v_0$  varies in the logarithm scale of base 10. Top 6 lines represent the true variables  $\mathbf{x}_{1:6}$  and the bottom 3 lines represent the maximum, median and minimum among the noise variables  $\mathbf{x}_{7:40}$ .

bootstrap which allows us to deal with just a subset of variables of size smaller than  $p$ .

## 5.4 A real example

For TFI, a company that owns some of the world's most well-known brands like Burger King and Arby's, decisions on where to open new restaurants are crucial. It usually takes a big investment of both time and capital at the beginning to set up a new restaurant. If a wrong location is chosen, likely the restaurant will soon be closed and all the initial investment will be lost. TFI hosted a prediction competition on Kaggle<sup>1</sup>, where the goal is to build a mathematical model to predict the revenue of a restaurant based on a set of demographic, real estate, and commercial information. The data contains 137 restaurants in the training set and 1000 restaurants in the test set. Features include the Open Date,

<sup>1</sup><https://www.kaggle.com/c/restaurant-revenue-prediction>



	$x_j \in \text{Signal}$	$x_j \in \text{Noise}$
BBEM (BIC)	2.99	0.24
BBEM ( $v_0 = 0.03$ )	2.96	0.27
EMVS	2.97	0.29
Oracle	3	0

Table 4: A large- $p$  small- $n$  example. The table shows the average number of signal and noise variables being selected out of 100 iterations. In BBEM,  $v_0$  is either chosen by BIC or fixed at 0.03. EMVS is the algorithm proposed by Ročková and George (2014).

City, City Group, Restaurant Type, and three categories of obfuscated data (P1-P37, numeric): demographic data, real estate data, and commercial data. The response is the transformed restaurant revenue in a given year.

We first transform the “Open Date” to a numeric feature called “Year Since 1900” and merge the “City” column into the “City Group” column which now contains four categories: Istanbul, Izmir, Ankara, and others (small cities). Then we create dummy variables for the categorical features like “City Group” and “Restaurant Type” and keep all the obfuscated numeric columns P1-P37. The final training set has 43 features and 137 samples.

After standardizing the data, we fix  $v_1$  at 100 and tune  $v_0$  from  $10^{-4.5}$  to  $10^{-0.5}$  for the BBEM algorithm, where each bootstrap sample uses  $L = 15$  variables, and the total number of replicates is  $K = 300$ . The path-plot of selection frequency for important features is shown in Figure 5.4. It is not surprising that “City Group”, “Years Since 1900” and “Restaurant Type” are important predictors for the revenue. Quite a few obfuscated features are also selected as important predictors. Although we do not know their meanings, they should provide valuable information for TFI to choose their next restaurant’s location.

Since the evaluation metric for this specific competition is based on the rooted mean square error (RMSE), we use the same metric in our 5-fold cross-validation. We tuned  $v_0$  from the set  $\{0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01\}$ , and found  $v_0 = 0.002$  has the smallest RMSE score. Then we fix  $v_0$  at 0.002, and re-run BBEM on the whole training data. Let  $\mathbf{m}$  denote the averaged posterior mean of  $\beta$  from  $L$  bootstrap iterations, and  $\gamma$  the averaged selection frequency for  $p$  variables. We then use  $\mathbf{m} * \gamma$  (where  $*$  denotes element-wise product) for prediction in the same spirit as the Bayesian model averaging. Our final Kaggle score is 1989762.52, which outperforms the random forest benchmark

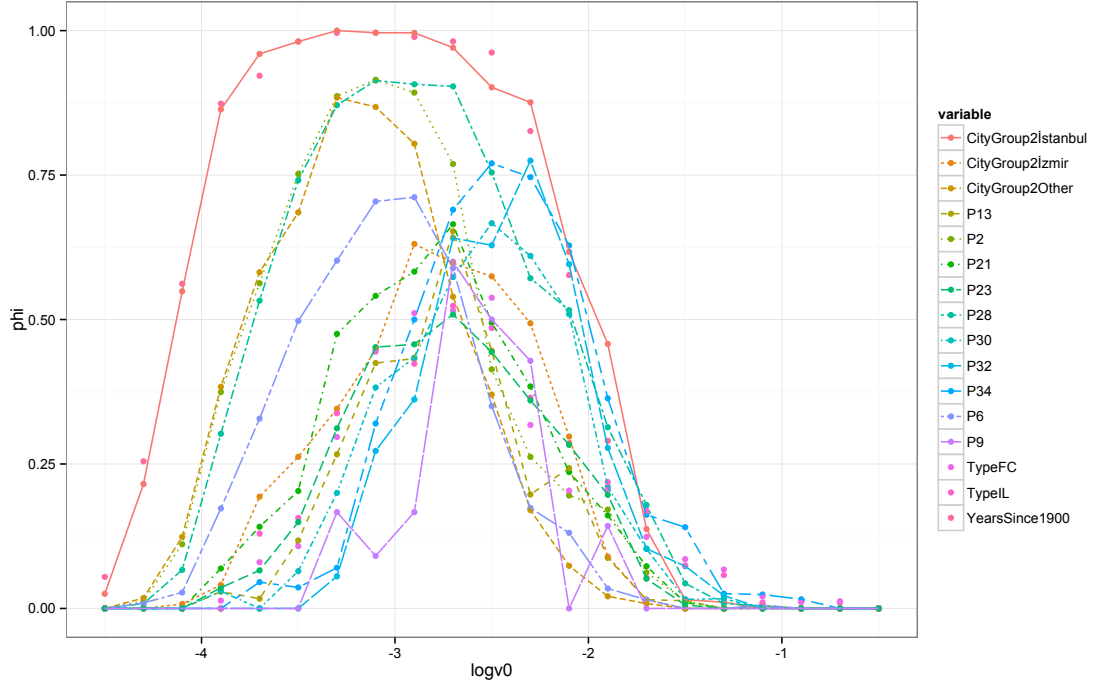


Figure 2: Restaurant data. The path plot of selection frequency when  $v_0$  varies in the logarithm scale of base 10. Only a subset of variables with high selection frequencies are displayed.

(RMSE=1998014.94) provided by Kaggle<sup>2</sup>. It is impressive for BBEM to outperform random forest considering that BBEM does not use any nonlinear features but random forest does.

## 6 Further Discussion

Variable selection is an important problem in modern statistics. In this paper, we study the Bayesian approach to variable selection in the context of multiple linear regression. We proposed an EM algorithm that returns the MAP estimate of the set of relevant variables. The algorithm can be operated very efficiently and therefore can scale up with big data. In

---

<sup>2</sup>At Kaggle, each team can submit their prediction and see the corresponding performance on the test data many times, so one can easily obtain a good score by keep tweaking the model to overfit the test data. For this reason, we did not compare our result with those “low” scores on the leaderboard provided by individual teams.

addition, we have shown that the MAP estimate from our algorithm provides a consistent estimator of the true variable set even when the model dimension diverges with the sample size. Further, we propose an ensemble version of our EM algorithm based on Bayesian bootstrap, which, as demonstrated via real and simulated examples, can substantially increase accuracy while maintaining the computation efficiency.

Although we restrict our discussion for the linear model, the two algorithm we proposed can be easily extended to other generalized linear models by using latent variables (Polson et al., 2013), an interesting topic for future research.

## Appendix: Proof of theorem 3.1

*Proof.* Recall the EM algorithm returns

$$\hat{\gamma}_{nj} = 1, \quad \text{if} \quad \mathbb{E}_{\beta|\Theta^{(t)}, \mathbf{y}}[\beta_j^2] > r_n,$$

where the threshold

$$r_n = \frac{\hat{\sigma}_n^2}{1/v_0 - 1/v_1} \left( \log \frac{v_1}{v_0} - 2 \log \frac{\hat{\theta}_n}{1 - \hat{\theta}_n} \right) = O(n^{-r_0} \log n)$$

and the conditional second moment of  $\beta_j$  is equal to  $m_j^2 + \hat{\sigma}_n^2 V_{jj}$  with

$$\begin{aligned} \mathbf{m} &= (\mathbf{X}_n^T \mathbf{X}_n + D^{-1})^{-1} \mathbf{X}_n^T (\mathbf{X}_n \beta_n^* + \mathbf{e}_n) \\ &= \beta^* - (\mathbf{X}_n^T \mathbf{X}_n + D^{-1})^{-1} D^{-1} \beta_n^* + (\mathbf{X}_n^T \mathbf{X}_n + D^{-1})^{-1} \mathbf{X}_n^T \mathbf{e}_n \\ &= \beta^* - \mathbf{b}_n + \mathbf{W}_n \\ \mathbf{V} &= (\mathbf{X}_n^T \mathbf{X}_n + D^{-1})^{-1}, \quad D^{-1} = \text{diag} \left( \frac{1 - \hat{\gamma}_{nj}}{v_0} + \frac{\hat{\gamma}_{nj}}{v_1} \right). \end{aligned}$$

Here we represent the posterior mean of  $\beta$  as three separate terms: the true coefficient vector  $\beta_n^*$ , the bias term  $\mathbf{b}_n$  and the random error term  $\mathbf{W}_n$ . So the event  $\{\hat{\gamma}_n = \gamma_n^*\}$  is equivalent to

$$\left\{ \min_{j: \gamma_{nj}^* = 1} m_j^2 + \hat{\sigma}_n^2 V_{jj} > r_n \right\} \cap \left\{ \max_{j: \gamma_{nj}^* = 0} m_j^2 + \hat{\sigma}_n^2 V_{jj} < r_n \right\}. \quad (21)$$

First we prove the following results that quantify  $m_j^2$  and  $V_{jj}$ .

(R1)  $V_{jj}$  is upper bounded by the largest eigenvalue of  $\mathbf{V}$ ,

$$V_{jj} \leq \frac{1}{\lambda_{n1} + 1/v_1} = O(n^{-\eta_1}) \prec O(n^{-r_0} \log n) = r_n, \quad (22)$$

where for two sequences  $\{a_n\}$  and  $\{b_n\}$ , we write  $a_n \prec b_n$  if  $a_n/b_n \rightarrow 0$ .

(R2) The bias term  $\mathbf{b}_n$  is bounded by

$$\begin{aligned} \max_j |b_{nj}| \leq \|\mathbf{b}_n\|_2 &\leq \|(\mathbf{X}_n^T \mathbf{X}_n + D^{-1})^{-1}\|_2 \cdot \|D^{-1} \boldsymbol{\beta}_n^*\|_2 \\ &\leq \frac{1/v_0}{\lambda_{n1} + 1/v_1} \|\boldsymbol{\beta}_n^*\|_2 = O(n^{r_0 - \eta_1 + \eta_2}). \end{aligned} \quad (23)$$

When  $r_0 < 2(\eta_1 - \eta_2)/3$ ,  $\max_j |b_{nj}|^2 \prec O(n^{-r_0} \log n) = r_n$ .

The matrix  $L_2$  norm is defined as  $\|A\|_2 = \sup_{\|v\|=1} \|Av\|_2$ , which is equal to its largest eigenvalue (singular value) when  $A$  is symmetric (non-symmetric).

(R3) Note that  $\mathbf{W}_n$  is not a Gaussian random vector due to the dependence between  $D$  and  $\mathbf{e}_n$ , but it can be rewritten as

$$\mathbf{W}_n = (\mathbf{X}_n^T \mathbf{X}_n + D^{-1})^{-1} (\mathbf{X}_n^T \mathbf{X}_n) (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{e}_n = A \tilde{\mathbf{W}}_n.$$

where  $A = (\mathbf{X}_n^T \mathbf{X}_n + D^{-1})^{-1} (\mathbf{X}_n^T \mathbf{X}_n)$  and  $\tilde{\mathbf{W}}_n = (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{e}_n$ . Since  $A$  is a matrix with norm bounded by 1, we have

$$\max_j |W_{nj}| \leq \|A\|_\infty \max_j |\tilde{W}_{nj}| \leq \sqrt{p} \|A\|_2 \max_j |\tilde{W}_{nj}| \leq \sqrt{p} \max_j |\tilde{W}_{nj}|.$$

(R4)  $\tilde{\mathbf{W}}_n = (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{e}_n$  is a Gaussian random vector with covariance  $\sigma^2 (\mathbf{X}_n^T \mathbf{X}_n)^{-1}$  and mean  $\mathbf{0}$ . So the variance for  $W_{nj}$  is upper bounded by  $\sigma^2 \lambda_{n1}^{-1}$ .

Recall the tail bound for Gaussian variables: for any  $Z \sim \mathcal{N}(0, \tau^2)$ ,

$$\mathbb{P}(|Z| > t) = \mathbb{P}(|Z|/\tau > t/\tau) \leq \frac{\tau}{t} e^{-\frac{t^2}{2\tau^2}}.$$

With Result (R3) and Bonferroni's inequality, we can find a constant  $M > 0$  such that

$$\begin{aligned} \mathbb{P}(\max_j |W_{nj}| > \sqrt{r_n}) &\leq \mathbb{P}(\max_j |\tilde{W}_{nj}| > \sqrt{r_n/p}) \\ &\leq p \cdot \mathbb{P}(|\tilde{W}_{nj}| > \sqrt{r_n/p}) \\ &\leq \frac{p\sqrt{p}\sigma}{\sqrt{r_n \lambda_{n1}}} e^{-\frac{r_n \lambda_{n1}}{2p\sigma^2}} = O(e^{-Mn^{\eta_1 - r_0 - \alpha}}), \end{aligned}$$

which goes to 0 when  $r_0 < \eta_1 - \alpha$ . So with probability going to 1,  $\max_j |W_{nj}|$  is upper bounded by  $\sqrt{r_n}$ .

(R5) When  $1 - \eta_3 < r_0$ ,  $\min_{j: \gamma_{nj}^* = 1} |\beta_{nj}^*|^2 \sim n^{\eta_3 - 1} \succ O(n^{-r_0} \log n) = r_n$ .

Now we prove (21). Given  $1 - \eta_3 < r_0 < \min\{\eta_1 - \alpha, 2(\eta_1 - \eta_2)/3\}$ , we have

$$\begin{aligned} \mathbb{P}\left(\max_{j:\gamma_{nj}^*=0} (m_j^2 + \hat{\sigma}_n^2 V_{jj}) > r_n\right) &\leq \mathbb{P}\left(\left(\max_j |b_{nj}| + \max_j |W_{nj}|\right)^2 + \hat{\sigma}_n^2 \max_j V_{jj} > r_n\right) \\ &\leq \mathbb{P}\left(\max_j |W_{nj}| > \sqrt{r_n}\right) = O(e^{-Mn^{\eta_1-r_0-\alpha}}), \\ \mathbb{P}\left(\min_{j:\gamma_{nj}^*=1} (m_j^2 + \hat{\sigma}_n^2 V_{jj}) < r_n\right) &\leq \mathbb{P}\left(\min_{j:\gamma_{nj}^*=1} |\beta_{nj}^*|^2 - \left(\max_j |b_{nj}| + \max_j |W_{nj}|\right)^2 < r_n\right) \\ &\leq \mathbb{P}\left(\max_j |W_{nj}| > \sqrt{r_n}\right) = O(e^{-Mn^{\eta_1-r_0-\alpha}}). \end{aligned}$$

So (21) holds with probability  $1 - O(e^{-Mn^{\eta_1-r_0-\alpha}}) \rightarrow 1$ .  $\square$

## References

- Breiman, L. (1996), “Bagging Predictors,” *Machine Learning*, 24, 123–140.
- (2001), “Random Forests,” *Machine Learning*, 45, 5–32.
- Clyde, M. A. and Lee, H. K. H. (2001), “Bagging and Bayesian Bootstrap,” in *Artificial Intelligence and Statistics*, eds. Richardson, T. and Jaakkola, T., pp. 169–174.
- Fan, J. and Li, R. (2001), “Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties,” *Journal of the American Statistical Association*, 96, 1348–1360.
- George, E. I. and McCulloch, R. E. (1993), “Variable Selection via Gibbs Sampling,” *Journal of the American Statistical Association*, 88, 881–889.
- Laurent, B. and Massart, P. (2000), “Adaptive Estimation of A Quadratic Functional By Model Selection,” *The Annals of Statistics*, 28, 1302–1338.
- Mathai, A. and Provost, S. (1992), *Quadratic Forms in Random Variables*, Statistics: A Series of Textbooks and Monographs, Taylor & Francis.
- Meinshausen, N. and Bühlmann, P. (2010), “Stability Selection,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 417–473.
- Mitchell, T. J. and Beauchamp, J. J. (1988), “Bayesian Variable Selection in Linear Regression,” *Journal of the American Statistical Association*, 83, 1023–1032.

- O'Hara, R. B. and Sillanpää, M. J. (2009), "A Review of Bayesian Variable Selection Methods: What, How and Which," *Bayesian Analysis*, 4, 85–118.
- Polson, N. G., Scott, J. G., and Windle, J. (2013), "Bayesian Inference for Logistic Models Using Polya-Gamma Latent Variables," *Journal of the American Statistical Association*, 108, 1339–1349.
- Ročková, V. and George, E. I. (2014), "EMVS: The EM Approach to Bayesian Variable Selection," *Journal of the American Statistical Association*, 109, 828–847.
- Rubin, D. B. (1981), "The Bayesian Bootstrap," *The Annals of Statistics*, 9, 130–134.
- Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58, 267–288.
- Wang, S., Nan, B., Rosset, S., and Zhu, J. (2011), "Random Lasso," *The Annals of Applied Statistics*, 5, 468–485.
- Xin, L. and Zhu, M. (2012), "Stochastic Stepwise Ensembles for Variable Selection," *Journal of Computational and Graphical Statistics*, 21, 275–294.
- Zhu, M. and Chipman, H. A. (2006), "Darwinian Evolution in Parallel Universes: A Parallel Genetic Algorithm for Variable Selection," *Technometrics*, 48, 491–502.